Framework para el procesamiento lingüístico de artículos científicos. Caso de estudio: Universidad Nacional de Chilecito

Jose Texier[1], Jusmeidy Zambrano[1], Emmanuel Frati[1]
[1] Universidad Nacional de Chilecito, Chilecito-La Rioja, Argentina
{jtexier, jzambrano, fefrati}@undec.edu.ar

Resumen. El crecimiento de los datos producto del Acceso Abierto a las publicaciones académicas y científicas han generado estudios que posibilitan la interrelación de áreas como la lingüística y la computación en, por ejemplo, la extracción automática de datos para la creación de modelos teóricos, reconocimiento de estructuras para validación, etc. Este trabajo tiene como objetivo describir un framework elaborado a partir de la Rhetorical Structure Theory (RST) con el lenguaje Python a un corpus de 42 artículos científicos en español de la Universidad Nacional de Chilecito. El análisis se hizo sobre la base de un diseño modular informático y el trabajo manual de un lingüista experto, proceso que sirve para calibrar la propuesta y, luego, ser aplicada a un corpus mayor. Las variables que se consideraron fueron el título, resumen y palabras clave de los 42 artículos y la estructura canónica de un resumen científico (introducción, método, resultados y discusión -IMRD). Los resultados muestran que existen discrepancias en la frecuencia de ciertos elementos en los textos, pero a su vez, denotan coincidencias interesantes para este tipo de análisis textuales.

Palabras clave: reconocimiento de texto, retórica, Python, artículos científicos, Universidad Nacional de Chilecito, RST, IMRD.

1 Introducción

El auge de la tecnología ha transformado la investigación científica en el siglo XX [1], [2]. Con la llegada de internet y, sobre todo, con el movimiento de Acceso Abierto la producción académica y científica de la mayoría de las instituciones (por ejemplo: universidades, colegios, institutos, etc.) ha estado sujeta a controversias y acuerdos respecto de su visibilidad, preservación, difusión y uso. La tendencia actual es que la información académica y científica sea accesible sin restricciones legales y técnicas [3], [4], de esta manera se posibilita para algunas ciencias interdisciplinarias la elaboración de modelos computacionales que reproduzcan uno o más aspectos del lenguaje humano. Estos modelos en el campo de la lingüística computacional se pueden lograr gracias a teorías lingüísticas; análisis morfológicos y sintácticos de textos en un idioma y contexto determinado [5].

En relación con esta disposición de grandes de datos textuales, en este artículo se

centra en describir un *framework* para el análisis lingüístico de artículos científicos visibles en fuentes de Acceso Abierto, en función de la producción científica (artículos en español) de la Universidad Nacional de Chilecito (UndeC). De esta producción se obtienen los metadatos principales: título, palabras clave y resumen, autores y filiación. El trabajo se realizó bajo un enfoque lingüístico basado en Rhetorical Structure Theory (RST) y con el lenguaje de programación Python y el paquete NLTK [6]. Python (a pesar de no contar con rapidez de cómputo) ofrece diversas ventajas relacionadas con la flexibilidad, curva de aprendizaje rápida, funcionalidad dada por las librerías para el análisis de texto (extensible) y una sintaxis y semántica transparente [6]–[8].

Este trabajo aporta otro punto de análisis vinculado con la producción científica y académica de las instituciones y, está en consonancia con las prioridades de Argentina respecto de la preservación y difusión del conocimiento científico. Las funciones de la Universidad Pública [9] y la promulgación de la Ley 26.899 (*Creación de Repositorios Digitales Institucionales de Acceso Abierto, Propios o Compartidos)* sostienen las bases para que el conocimiento producido al interior de las instituciones públicas se difunda a la comunidad en general. Por tanto, los sistemas informáticos (Repositorios Institucionales y/o Bibliotecas Digitales) de muchas instituciones que permiten preservar y difundir la producción, aportan una gran cantidad de datos estructurados que sirven de base para desarrollar otras propuestas.

El artículo presenta en una segunda sección, un marco teórico que da cuenta, *grosso modo*, de algunos estudios previos que sustentan la propuesta realizada; la siguiente sección describe el *framework* desarrollado; luego, se analizan y discuten los resultados obtenidos y, en la última sección se plantean algunas conclusiones y sugerencias para trabajos futuros.

2 Marco de Referencia

La enorme cantidad de información digital disponible en áreas de conocimiento ha facilitado el desarrollo de propuestas vinculadas con: la extracción automática de textos [10], la creación de sistemas expertos que respalden la labor de especialistas en un área dada [11], los análisis de grandes corpus para generación de modelos teóricos, etc. [5].

Al tomar como punto de partida la extracción de textos para evaluar los conceptos y/o términos presentes en ellos, se pueden encontrar diversos enfoques que lo hacen de forma automática [10], [12]: i) lingüístico, ii) estadístico, iii) aprendizaje automático, y iv) métodos híbridos. De manera general, el enfoque lingüístico intenta filtrar mediante patrones de información. Los enfoques estadísticos, usan un número diferente de medidas y distribuciones estadísticas. Los sistemas de aprendizaje automático usan datos para aprender rasgos que sean útiles y relevantes para el reconocimiento [10].

En este trabajo se utilizó un enfoque mixto tomando como inicio el enfoque lingüístico según Acosta [10] al usar una teoría discursiva y luego herramientas de aprendizaje automático a través del paquete NLTK de Python [6]. El corpus se conformó con resúmenes en español de los artículos especializados en revistas arbitradas cuyos autores tuvieran como filiación institucional la UNdeC. Se consideró el resumen (*abstract*) porque es el género textual utilizado con asiduidad en la comunicación científica que da cuenta, en un número breve de palabras, los objetivos, marcos, metodologías, resultados y discusiones en un tema [13].

Para el estudio se vinculan las Ciencias de Computación y la Lingüística, puesto que ambas toman como objeto de estudio el lenguaje, y, en el caso específico, la lingüística computacional toma el Procesamiento de Lenguaje Natural (PLN) para habilitar a las computadoras en la tarea de procesar y entender el texto [14], [15]. Por ello, la propuesta se desarrolló sobre la base teórica del Procesamiento del Lenguaje Natural y la teoría discursiva Rhetorical Structure Theory (RST) de Mann y Thompson [16], que se ha empleado en diversas aplicaciones, como generación de texto, resumen automático, traducción automática, extracción de información, etc. [17]. La mayor parte de estas aplicaciones se han centrado en las lenguas: inglés, el japonés y el portugués [17]. El modelo usado para los resúmenes del corpus seleccionado fue el modelo de Swales [18], es decir, la estructura retórica: Introducción, Método, Resultados y Discusión (IMRD), elementos que se identificaron en el *framework* desarrollado.

Para el procesamiento del texto de los resúmenes, se analizaron cada una de las oraciones siguiendo la técnica de marcado del discurso (*Part-of-speech* - POS), que consiste en reconocer las entidades nombradas y extracción de información [19], [20]. El proceso se realizó en forma automatizada, a través de la asignación de una etiqueta (*tag* en inglés) de la categoría gramatical en cada palabra. También se usaron técnicas para identificar la raíz (o *stemmer*) de un conjunto de palabras similares, este proceso es conocido como Lematización, es decir, encontrar la raíz léxica de las palabras. Cuando un proceso de lematización no es posible de realizar, se recurre a un proceso de truncado (en inglés es *Chunker*), cuya finalidad es aproximar lo más posible las palabras a su raíz léxica. El etiquetado se basó en las propuestas por el grupo EAGLES (*Expert Advisory Group on Language Engineering Standards*) [21].

Al colocar en contexto teórico la propuesta desarrollada, se realizó un relevamiento de artículos que han trabajado de manera similar el *framework* propuesto, clasificado en tres grandes temas: análisis de los artículos científicos, generación de resúmenes automáticos/reconocimiento de términos,y análisis retóricos/contextos definitorios.

En cuanto al análisis de la estructura de los artículos científicos se encuentran tres trabajos, uno presenta un método para caracterizar el enfoque (contribución principal), dominio y técnicas/herramientas usadas en el artículo a partir de un matching con patrones semánticos usando el aprendizaje de bootstrapping [22]. El otro trabajo analiza los artículos científicos biomédicos a partir de tres esquemas: nombre de las secciones, zonas argumentativas y core de conceptos científicos, usando machine learning [23]. El tercer trabajo realizado en el 2008, se centra en la estructura retórica de los resúmenes científicos aplicando minería de texto gracias a

un corpus de Medline aplicando machine learning, identificando cuatro secciones objetivos, métodos, resultados y conclusiones [24].

El tema de generación de resúmenes automáticos es relevante porque representa el proceso inverso al *framework* propuesto. Teufel y Moens [25] realizan una investigación sobre el tema, sin embargo, se encuentran tres trabajos que extraen información en forma automatizada usando diferentes temas, pero utilizando principios de la generación de resúmenes automáticos [10], [26], [27]. El trabajo más importante de los tres es el laboratorio en línea presentado por Torres, ya que cuentan con un sistema de cuatro módulos: preprocesado, etiquetado POS, identificación de entidades nombradas (abreviatura en inglés NER) y *parse tree* [27].

En el tercer gran tema, se encuentran trabajos que se caracterizan por el uso retórico en los análisis de producciones científicas [24], [28], [29] y desde el punto de contextos definitorios [30], [31]. Hirohata et al. [24] identifica cuatro secciones en los abstract de las producciones científicas: objetivos, métodos, resultados y conclusiones. Prabhakaran et al. [28] hacen un análisis de 2.4 millones de abstracts de la Web of Science desde 1991 al 2010 y pueden determinar el crecimiento y/o declive de tópicos retóricos de ese corpus. En el tercer trabajo sobre análisis retóricos, los autores hacen un chequeo de las relaciones retóricas a partir de la RST (en Vasco) con el corpus de la revista médica de Bilbao [29]. En cuanto a la extracción de fragmentos textuales que contengan definiciones vinculadas por predicaciones verbales (contextos definitorios), se destaca el trabajo de Aguilar et. al. [30], porque definen un método enfocado a la biomedicina. El trabajo de Sierra [31], presenta un estado del arte en este contexto para entenderlo y estudiar la posibilidad de incorporar estas funcionalidades en una ampliación del *framework* desarrollado.

Actualmente, existen muchas herramientas que realizan PLN con virtudes y carencias según el objetivo deseado. A continuación se presenta un relevamiento de las herramientas de software encontradas:

- AnCora es un corpus del catalán (AnCora-CA) y del español (AnCora-ES) con diferentes niveles de anotación, http://clic.ub.edu/corpus/es.
- TreeBank, denominado también como Corpus parseado o más ampliamente Penn Treebank, http://www.comp.leeds.ac.uk/amalgam/tagsets/upenn.html.
- El RST Spanish Treebank, corpus en línea de textos especializados en español con relaciones discursivas de la RST de Mann y Thompson [16].
- Laboratorio en línea para el procesamiento automático de documentos [27].
- Stanford's Natural Language Processing, conjunto de herramientas de PLN, para documentos de texto los cuales pueden estar en diferentes idiomas [32]
- Natural Language Toolkit (NLTK), cuenta con un conjunto de bibliotecas de procesamiento de textos para la clasificación, simbolización, derivado, etiquetado, análisis sintáctico y semántico [6].
- TreeTagger, http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/.
- FreeLing, http://nlp.lsi.upc.edu/freeling/.
- Weka, http://www.cs.waikato.ac.nz/ml/weka/.
- Wagsoft Linguistic Software, http://www.wagsoft.com/software.html.

El *framework* desarrollado se ubica dentro del primer gran tema, análisis de la estructura de los artículos científicos, y se destaca que existe una implementación a

medida, a pesar de la existencia de una gran cantidad de herramientas, ya que la solución se posiciona en la teoría discursiva RST según la estructura retórica IMRD de Swales para resúmenes científicos.

3 *Framework* en Prueba

Las diferentes herramientas de software observadas y los trabajos relevados dan cuenta de que no existe un sistema específico que permita el análisis lingüístico de los resúmenes de artículos científicos, a partir del modelo de Swales [18]. De esta manera se puede verificar automáticamente que los resúmenes cumplan con la estructura textual canónica de "Introducción, Método, Resultados y Discusión - IMRD". El *framework* desarrollado se analizó sobre la base de esa estructura canónica y con tecnologías de licencia abierta. En fases posteriores, el sistema evolucionará luego de un proceso de revisión, incorporando nuevas funcionalidades y analizando corpus más grandes. El sistema se realizó bajo un enfoque lingüístico basado en RST y con el lenguaje de programación Python. El *framework* está estructurado en seis etapas (ver Fig. 1):

1. Recolección de la producción científica de la UNdeC
2. Selección de los artículos científicos en revistas de investigación en idioma español.
3. Extracción de los metadatos necesarios para el análisis.
4. Análisis de palabras clave.
5. Análisis de los títulos.
6. Análisis de los resúmenes.

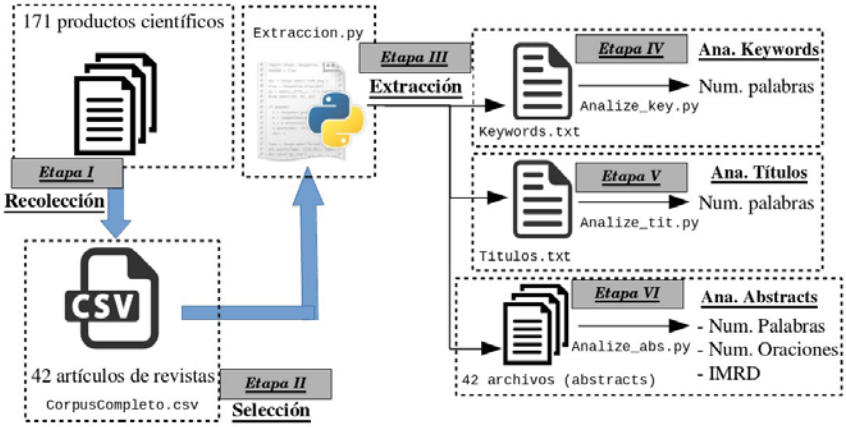

Fig. 1. *Framework* desarrollado.

3.1 Recolección

El corpus de la UNdeC, tomado el 30 de noviembre de 2016, se obtuvo de las siguientes fuentes: Google Scholar, RedALyC, SciELO y Scopus. Se realizó una

depuración de los resultados obtenidos con un total de 171 productos académicos y científicos con la afiliación de la Universidad Nacional de Chilecito (con al menos un autor) de esas cuatro fuentes. Los productos obtenidos tienen la siguiente tipología:

- Artículos en revistas: 100
- Conferencias: 53
- Libros: 3
- Reportes: 7
- Tesis: 8

Toda esta producción se incorporó en OpenRefine [33], herramienta que ofrece funcionalidades adicionales a los gestores de hojas de cálculo como LibreOffice Calc o Excel. Estos archivos y el proceso detallado de la investigación realizada se encuentran en un proyecto GitHub [34], el cual se puede consultar sin restricciones de acceso, legales ni técnicas.

3.2 Selección

De los 171 productos de la UNdeC, se seleccionaron los artículos de revistas en español. El total fue de 42 artículos extraídos en:

- Google Scholar, 24 artículos.
- RedALyC, 3 artículos.
- SciELO, 6 artículos.
- Scopus, 9 artículos.

Los metadatos de estos 42 artículos se almacenaron en un archivo CSV (*comma-separated values*) para adecuar la información a las siguientes etapas y poder adaptarlo a otros módulos en un futuro (Big Data por ejemplo).

3.3 Extracción

En esta etapa se realizó el proceso de extracción de metadatos del archivo CSV que proviene de la etapa II (`CorpusCompleto_UNdeC.csv`). En este proceso se usó el lenguaje de programación Python a través de un programa llamado `extraccion.py`. Esta etapa tiene como salida cuarenta y cuatro (44) archivos, que sirven de entrada para las etapas IV, V y VI. Los archivos son:

- Para la etapa IV se tiene el archivo `Keywords.txt`.
- En la etapa V se usó el archivo `Titulos.txt`.
- Para el análisis de los abstract (etapa VI) se usaron los 42 archivos que representan los 42 resúmenes del corpus seleccionado.

3.4 Análisis de palabras clave

De la etapa anterior se obtuvo el archivo "`Keywords.txt`", el cual sirve de entrada

a un programa en Python llamado "`AnalyzeKeywords.py`", donde se cuenta la cantidad de palabras, es decir, esta etapa consiste en totalizar las palabras clave que tienen los 42 artículos seleccionados. Al estar separada esta etapa en el *framework*, al igual que en las etapas restantes, permite incorporar a futuro cualquier tipo de proceso de acuerdo con los requerimientos deseados.

3.5 Análisis de títulos

En la etapa III se genero el archivo "`Titulos.txt`", el cual sirve de entrada al programa de Python "`AnalyzeTitle.py`", donde se totaliza la cantidad de palabras, es decir, el proceso es similar a la etapa IV.

3.6 Análisis de resúmenes

A diferencia de las etapas IV y V, en esta etapa se realizó el análisis lingüístico de acuerdo con los lineamientos establecidos, a partir de los 42 archivos generados en la etapa III que contienen los 42 resúmenes del corpus seleccionado, el proceso se realiza en el programa de Python llamado "`AnalyzeAbstract.py`". En esta etapa se aprecia la modularidad diseñada en el *framework*, ya que se pueden realizar diversos procesos de forma independiente. Los resultados de esta fase se encuentran en la carpeta "`Analisis_Linguistico`" del proyecto GitHub [34]. En esta sección se realizan las siguientes actividades:

- Etiquetado de oraciones: se realizó a través de Python y el paquete Stanford CoreNLP [32], que tiene la posibilidad de hacer el etiquetado para el idioma español.
- Extracción de verbos: la nomenclatura para el etiquetado es la de EAGLE [21], que clasifica los verbos de tres maneras en forma general: principal (`vm`), auxiliar (`va`) y semiauxiliar (`vs`). Luego del etiquetado se extrajeron todas las palabras (o tokens) que tienen la clasificación de verbo.
- Identificación de verbos: luego de extraer los verbos, se busca el lema o lexema (su forma canónica) usando los procesos de truncamiento (*Stem*), que luego se comparan con un listado de verbos clasificados por los autores en las categorías IMRD, de acuerdo con la estructura retórica de Swales [18]. Todo esto gracias al paquete NLTK de Python [6].

4 Resultados

El proceso realizado consistió en hacer uso del *framework* y compararlo con la revisión manual realizada por el experto lingüista. Este análisis manual del experto se encuentra en el proyecto GitHub [34] con el nombre "`experto.pdf`".

En cuanto a las diferentes etapas del sistema desarrollado produjeron los siguientes resultados:

- Para la sección de títulos, se encontraron 629 palabras, para un promedio de 14,98 palabras por título.
- En la sección de las palabras clave, se extrajeron 348 palabras, con un promedio de 8,49 por palabras clave.
- Para la sección de resúmenes, se obtuvieron 7.089 palabras, con un promedio de 172,90 por cada resumen.
- Un promedio de 6,42 oraciones por cada resumen de los artículos. El total de oraciones analizadas fue de 269, que pertencen al corpus de los 42 artículos analizados.

En cuanto al análisis del experto, se resume con la Tabla 1:

- Las diferencias entre el análisis manual del experto y el *framework* desarrollado no exceden el 20%. Por ejemplo, en el corpus de resúmenes analizados el experto consideró que el 78,57% tiene una "introducción" y el *framework* detectó 85,71%, con una diferencia de 7,14%, lo que indica que hubo mucha coincidencia entre ambos análisis.
- En el caso de sección "discusión" la diferencia es la más alta (19,04%) lo que demuestra que los análisis fueron desiguales, para el experto, en los resúmenes solo el 38,10% tiene discusión y para el *framework* el 57,14%.
- En un total de 13 artículos (30,95%), el *framework* y el experto coinciden que esos artículos tienen las cuatro secciones en sus resúmenes de acuerdo con Swales. En cuanto a coincidencias en solo tres aspectos son 18 artículos (42,86%), 7 artículos (16,67%) en dos y 4 artículos (9,52%) en uno.

Tabla 1. Resultados del *framework* y del experto

	EXPERTO	*FRAMEWORK*	VARIACIÓN-DIFERENCIA
Introducción	78,57%	85,71%	7,14%
Método	78,57%	90,48%	11,91%
Resultados	59,52%	73,81%	14,29%
Discusión	38,10%	57,14%	19,04%

5 Trabajos Futuros

Esta investigación puede servir de base para realizar los siguientes trabajos:

- Un análisis de términos de acuerdo con los tesauros de la UNESCO y de la OECD para obtener una lista que posibilite un enfoque temático por áreas y/o disciplinas.
- Una implementación con todo el corpus de la UNdeC, ya que este trabajo se realizó como primera fase con un corpus de prueba de 42 artículos, verificados manualmente y obtenidos por el *framework*.

- El *framework* estará evolucionando para seguir con el análisis de todos los artículos de las conferencias del Latin American and Caribbean Consortium of Engineering Institutions - LACCEI - (aproximadamente 1.500) y con los artículos de la Universidad Nacional de La Plata (aproximadamente unos 25.000 artículos). En esta fase, se agregaran otras funcionalidades para que forme parte de un proyecto bajo los principios del Big Data [35]–[37], ya que implica el manejo de grandes cantidades de datos (volumen), la diversidad de temáticas (variedad) y la adaptación de un flujo continuo de datos en tiempo casi real (velocidad).
- Realizar un análisis semántico de la producción a partir del desarrollo de una ontología.
- Replicar este trabajo con un corpus pequeño en inglés, que permita hacer una verificación manual de los resultados del sistema y contrastarlos con el análisis de un experto.
- Generar trabajos interdisciplinarios (lingüistas e informáticos) al interior de las universidades para la enseñanza de los géneros académicos y/o científicos que generen instancias automatizadas de los procesos de escritura de los textos.

6 Conclusiones

Este trabajo se planteó como objetivo central, el describir un *framework* elaborado a partir de la Rhetorical Structure Theory (RST) con el lenguaje Python para un corpus de 42 artículos científicos en español de la Universidad Nacional de Chilecito. El análisis se hizo sobre la base de un diseño modular informático con tecnologías abiertas y el trabajo manual de un lingüista experto, lo que permitió hacer una implementación a pequeña escala para poder verificar de forma manual y evolucionar a un sistema con un corpus mayor. Las variables que se consideraron fueron el título, resumen y palabras clave de los 42 artículos y la estructura canónica de un resumen científico. El diseño se basó en módulos para que en un futuro, los módulos trabajen de forma independiente aceptando cualquier tipo de documentos y bajo los principios de Big Data, de manera tal que se puedan adaptar los análisis de otros datos como tuits o noticias diarias, por ejemplo.

Los resultados muestran que existen discrepancias en la frecuencia de ciertos elementos en los textos, pero a su vez, denotan coincidencias interesantes para este tipo de análisis textuales. Al comparar los dos resultados sobre la retórica IMRD (*framework* y experto) realizados, se puede concluir que existe una gran similitud pero no suficiente, ya que siempre se desea alcanzar un ciento por ciento de coincidencia. Un total de 13 artículos (30,95%) según el *framework* y el experto coinciden que esos artículos tienen las cuatro secciones (IMRD) en sus resúmenes de acuerdo con Swales, por tal razón, hay que estudiar más en profundidad las temáticas y exigencias de cada revista para poder tomar posición clara ante el bajo porcentaje de coincidencia. Además, serán necesarios otros niveles de análisis para redefinir la estructura textual y poder adaptar el *framework* a la realidad discursiva de algunas disciplinas.

En la actualidad la competitividad académica se refleja en la visibilidad web de la institución, por lo que se hace necesario que la institución logre estándares que permitan mayor impacto, por ello, los resultados obtenidos en esta investigación pueden ayudar a generar políticas institucionales sobre este aspecto.

7 Referencias

[1] J. De Souza-Silva, J. Cheaz Peláez, and J. Calderón Romero, *La cuestión institucional, de la vulnerabilidad a la sostenibilidad institucional en el contexto del Cambio de Epoca*. Servicio Internacional para la Investigación Agrícola Nacional, 2001.

[2] M. Castells, *The Rise of the Network Society: The Information Age: Economy, Society, and Culture Volume I*, 2nd Edition with a New Preface. Wiley-Blackwell, 2009.

[3] J. Texier, "Los repositorios institucionales y las bibliotecas digitales: una somera revisión bibliográfica y su relación en la educación superior," presented at the 11th LACCEI, 2013, Cancun, Mexico, 2013, p. 9.

[4] P. Suber, "Ensuring open access for publicly funded research," *BMJ*, vol. 345, 2012.

[5] A. Domínguez Burgos, "Lingüistica computacinal: un esbozo," *Boletín de lingüística*, no. 18, pp. 104–119, 2002.

[6] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc., 2009.

[7] R. Ali, *Python: Python For Beginners Crash Course Master Python Programming Fast and Easy Today*, 1st ed. USA: CreateSpace Independent Publishing Platform, 2015.

[8] S. H. Edwards, D. S. Tilden, and A. Allevato, "Pythy: Improving the Introductory Python Programming Experience," in *Proceedings of the 45th ACM Technical Symposium on Computer Science Education*, New York, NY, USA, 2014, pp. 641–646.

[9] J.-I. Badell, C. Rovira, and M. Térmens, "Estudio de visibilidad web 2013 de los museos de Cataluña," *Ibersid: revista de sistemas de información y documentación*, vol. 8, 2014.

[10] O. L. Acosta, C. A. Aguilar, and T. Infante, "Reconocimiento de términos en español mediante la aplicación de un enfoque de comparación entre corpus" *Linguamática*.2015.

[11] G. Aquino and L. C. Lanzarini, "Keyword identification in spanish documents using neural networks," *Journal of Computer Science & Technology*, vol. 15, no. 2, Nov. 2015.

[12] J. A. Lossio-Ventura, C. Jonquet, M. Roche, and M. Teisseire, "Yet Another Ranking Function for Automatic Multiword Term Extraction," in *Advances in Natural Language Processing*, 2014, pp. 52–64.

[13] T. A. van Dijk, *La ciencia del texto: un enfoque interdisciplinario*. 1983.

[14] "Asociación Mexicana para el Procesamiento del Lenguaje Natural Main/Home Page." [Online]. Available: http://www.ampln.org/. [Accessed: 30-Apr-2017].

[15] M. Vallez and R. Pedraza, "El Procesamiento del Lenguaje Natural en la Recuperación de Información Textual y áreas afines," *Hipertext.net*, 2007.

[16] W. Mann and S. Thompson, "Rhetorical Structure Theory: Toward a functional theory of text organization," *Text-Interdisciplinary Journal for the Study of Discourse*,vol.8, 2009.

[17] I. D. Cunha, J.-M. Torres-Moreno, and G. Sierra, "Aplicaciones lingüísticas del análisis discursivo automático," *Comunicación Social en el Siglo XXI*, vol. II, 2011.

[18] J. Swales, *Genre Analysis: English in Academic and Research Settings*. Cambridge University Press, 1990.

[19] A. Ritter, S. Clark, Mausam, and O. Etzioni, "Named Entity Recognition in Tweets: An Experimental Study," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA, 2011, pp. 1524–1534.

[20] D. Ye, Z. Xing, J. Li, and N. Kapre, "Software-specific Part-of-speech Tagging: An Experimental Study on Stack Overflow," in *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, New York, NY, USA, 2016, pp. 1378–1385.

[21] "Expert Advisory Group on Language Engineering Standards (EAGLES)," 25-Mar-2016. [Online]. http://www.ilc.cnr.it/EAGLES96/home.html. [Accessed: 30-Apr-2017].

[22] S. Gupta and C. D. Manning, "Analyzing the Dynamics of Research by Extracting Key Aspects of Scientific Papers," *IJCNLP*, 2011.

[23] Y. Guo, A. Korhonen, M. Liakata, I. S. Karolinska, L. Sun, and U. Stenius, "Identifying the Information Structure of Scientific Abstracts: An Investigation of Three Different Schemes," in *2010 Workshop on Biomedical Natural Language Processing*, USA, 2010

[24] K. Hirohata, N. Okazaki, S. Ananiadou, M. Ishizuka, and M. I. Biocentre, "Identifying Sections in Scientific Abstracts using Conditional Random Fields," *IJCNLP*, 2008.

[25] S. Teufel and M. Moens, "Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status," *Computational Linguistics*, vol. 28, no. 4, pp. 409–445, Dec. 2002.

[26] P. Saint-Dizier, "Processing natural language arguments with theplatform," *Argument & Computation*, vol. 3, no. 1, pp. 49–82, Mar. 2012.

[27] J. López, C. Sánchez-Sánchez, and E. Villatoro-Tello, "Laboratorio en lınea para el procesamiento automático de documentos," 2014.

[28] V. Prabhakaran and O. Rambow, "Predicting Power Relations between Participants in Written Dialog from a Single Thread - P14-2056," 2014. [Online]. Available: http://www.aclweb.org/anthology/P14-2056. [Accessed: 30-Apr-2017].

[29] M. Iruskieta, "The RST Basque TreeBank: an online search interface to check rhetorical relations."

[30] C. Aguilar, O. Acosta, G. Sierra Martínez, S. Juárez, and T. Infante, "Extracción de contextos definitorios en el área biomedicina," *Extraction of Definitional Contexts from Biomedical Corpora*, Sep. 2016.

[31] G. Sierra, "Extracción de contextos definitorios en textos de especialidad a partir del reconocimiento de patrones lingüísticos," *Linguamática*, vol. 1, no. 2, 2009.

[32] "Stanford CoreNLP – Natural language software | Stanford CoreNLP." [Online]. Available: https://stanfordnlp.github.io/CoreNLP/index.html. [Accessed: 27-Jun-2017].

[33] OpenRefine, "OpenRefine," 2017. [Online]. http://openrefine.org/. [Accessed:Apr-2017].

[34] J. Texier and J. Zambrano, "Frameword-PC - GitHub." [Online]. Available: https://github.com/dantexier/Frameword-PC. [Accessed: 30-Apr-2017].

[35] V. Rajaraman, "Big data analytics," *Reson*, vol. 21, no. 8, pp. 695–716, Aug. 2016.

[36] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *International Journal of Information Management*, vol. 35, 2015.

[37] M. Tascón, "Introducción: Big Data. Pasado, presente y futuro," *Telos: Cuadernos de comunicación e innovación*, no. 95, pp. 47–50, 2013.